

Bug hunting with Apache Lucene

Uwe Schindler

Apache Lucene PMC & Apache Software Foundation Member

uschindler@apache.org

<http://www.thetaphi.de>, <http://blog.thetaphi.de>

 @ThetaPh1

SD DataSolutions GmbH, Wätjenstr. 49, 28213 Bremen, Germany

Tel: +49 421 40889785-0, <http://www.sd-datasolutions.de>

My Background

- **Committer** and **PMC member** of **Apache Lucene** and **Solr**
- main focus is on development of Lucene Core.
- Member of **Apache Software Foundation**
- Well known as *Generics and Sophisticated Backwards Compatibility Policeman*.
- Working as **consultant** and **software architect** at **SD DataSolutions GmbH** in Bremen, Germany.
- Maintaining PANGAEA (Publishing Network for Geoscientific & Environmental Data) the first portal that used Apache Lucene for Geographical SearchApache Lucene Core and Elasticsearch.



Apache Lucene Core is a high-performance, full-featured text search engine library **written entirely in Java.**

It is a technology suitable for nearly any application that requires full-text search, especially cross-platform.

Inverted Index

328

Index

- Canvas of life turned upside down, 68
"Carbonate of pork," 315
Carracci, the, 147
Casina di Banda, 119
Castelletto, 265
Cavagnago, 76
Cenere, Monte, narcissuses on, 228
Ceres, 161
Cerrea, 133
Chalk, Conté, the Italian for whom this was the one thing needful, 136
Chalk eggs, 43
Chamois, foot of, 283
Change, repudiation of desire for sudden, 186
— importance of, depends on the rate of introduction, 196
— either the circumstances or the sufferer will, 196
Changes, sweeping, to be felt hereafter as vibrations, 60
Cheapissimo, 165
Cheese and the *alpi*, 289
Cherries, 33, 35, 46
Chestnuts, 118
Chicory and seed onions, weary utterness in, 227
Children, subalpine, 301
— what becomes of the clever, 149
Chinese, the examination-ridden, 151
Chironico, 75
"Chow," 52
Church-going, subalpine, 303
Circulation of people like blood, 20
Ciseri, his picture at Locarno, 271
Civilisation, antiquity of Italian, 124
— stationary, of ants and bees, 195
Class distinction inevitable, 195
Classification only possible through sense of shock, 63
Clergy, our English and S. Michele
- Cocking, Wednesbury, 55, 305
Collects, unsympathetic priest bristling with, 111
Colleone, Medea, 231
Colma di San Giovanni, 163
Comba di Susa, 119
Comfort as a moral influence, 185
Comic song, the landlord's, 128
Common sense, the safest guide, 108
Consistent, who ever is? 153
Contradictory principles, there must be a harmonious fusing of, 152
Converting things by eating them, 153
Corpses, desiccated, at S. Michele, 97
Cousins, my, the lower animals, 60
Cows fighting in farmyard, 120
Cricco, 125
Cristoforo, S., church of, at Mesocco, 208
— — — — at Castello, 234
Crossing, efficacy of, 152
— unexpected results of, 55
— useless if too wide, 157
Crucifixion, fresco at Fusio, 140
Culture and priggishness, 141
— a mode of concealing weakness, 192
Current feeling, the safest guide, 108
Cutlets, burnt, and the waiter, 144
- Dalpe, 38
Dante a humbug, 156
Darwin, Charles, no place for meeting, 69
Darwin, Erasmus, 23, 133
Dazio, Signor Pietro, of Fusio, 140
Death, no man can die to himself, 277
Deceit a necessary alloy of truth, 280

Index

329

- Department, good technique resembles, 156
Desire and power, 108
Development of power to know our own likes and dislikes, 22
Devil's Bridge, 23
Diatonic scale, and song of birds in New Zealand, 232
Dirt, eating a peck of moral, 71
Disgrazia and misfortune, 58
D'Israeli, Isaac, quotations from, 67
Dissenters all narrow-minded, 153
Distribution of plants and animals often inexplicable, 135
Diversion of mental images, 54
Doera, fresco at, 145, 221
Dogs, 156, 202, 260, 313
Doing, the only mode of learning, 151
Doors, how they open in time, 151
Doubt, "There lives more doubt in honest faith," 67
Downs, the South, like Monte Generoso, 230
Draughtsman, first business of a, 148
Drawing, the old manner of teaching, 150
Dream, my, at Lago di Cadagno, 82
Drunkenness and imagination, 46
Dunque, 133
Duso, Agostino, his fresco at Sta. Maria in Calanca, 225
- Earnestness, 142, 192
Eating, a mode of bigotry, 153
Echo at Graglia, 192
Edelweiss, 291
Electricity and Alpine roads, 60
Elephant brays a third, 233
"Elongated" honey, 293
Embryonic stages, the artist
- English priests and Italian, 106
— why introspective, 18
Equilibrium only attainable at the cost of progress, 195
Eritis, a panic concerning, 204
Eternal punishment, 111, 196
Eusebius, St., 178
Evolution and illusion, 43
— essence of, consists in not shocking too much, 110
Extreme, every, an absurdity, 153
- Faido, 22
Faith, doubt lives in honest, 67
— more assured in the days of spiritual Saturnalia, 68
— foundations of our system based on, 107, 277
— and reason, 108
— catholic, of protoplasm, 152
— a mode of impudence, 283
Falsehood turning to truth, 71
Famine prices at Locarno, 276
Feeling, current, the safest guide, 108
Fertile, rich and poor rarely fertile *inter se*, 195
Fires, how Italians manage their, 117
Fishmonger choosing a bloater, 23
Flats and sharps, a maze of metaphysical, 23
Fleet Street, beauties of, 19
Flowers, names of, 291
Fossil-soul, 234
Foundations of action lie deeper than reason, 107
— of a durable system laid on faith, 277
Francis, St., and Insurance Co.'s plate, 191
Friction, which prevents the unduly rapid growth of inventions, 60

Inverted Index

328

Index

- Canvas of life turned upside down, 68
"Carbonate of pork," 315
Carracci, the, 147
Casina di Banda, 119
Castelletto, 265
Cavagnago, 76
Cenere, Monte, narcissuses on, 228
Ceres, 161
Cerrea, 133
Chalk, Conté, the Italian for whom this was the one thing needful, 136
Chalk eggs, 43
Chamois, foot of, 283
Change, repudiation of desire for sudden, 186
— importance of, depends on the rate of introduction, 196
— either the circumstances or the sufferer will, 196
Changes, sweeping, to be felt hereafter as vibrations, 60
Cheapissimo, 165
Cheese and the *alpi*, 289
Cherries, 33, 35, 46
Chestnuts, 118
Chicory and seed onions, weary utterness in, 227
Children, subalpine, 301
— what becomes of the clever, 149
Chinese, the examination-ridden, 151
Chironico, 75
"Chow," 52
Church-going, subalpine, 303
Circulation of people like blood, 20
Ciseri, his picture at Locarno, 271
Civilisation, antiquity of Italian, 124
— stationary, of ants and bees, 195
Class distinction inevitable, 195
Classification only possible through sense of shock, 63
Clergy, our English and S. Michele
Cocking, Wednesbury, 55, 305
Collects, unsympathetic priest bristling with, 111
Colleone, Medea, 231
Colma di San Giovanni, 163
Comba di Susa, 119
Comfort as a moral influence, 185
Comic song, the landlord's, 128
Common sense, the safest guide, 108
Consistent, who ever is? 153
Contradictory principles, there must be a harmonious fusing of, 152
Converting things by eating them, 153
Corpses, desiccated, at S. Michele, 97
Cousins, my, the lower animals, 60
Cows fighting in farmyard, 120
Cricco, 125
Cristoforo, S., church of, at Mesocco, 208
— — — — at Castello, 234
Crossing, efficacy of, 152
— unexpected results of, 55
— useless if too wide, 157
Crucifixion, fresco at Fusio, 140
Culture and priggishness, 141
— a mode of concealing weakness, 192
Current feeling, the safest guide, 108
Cutlets, burnt, and the waiter, 144
Dalpe, 38
Dante a humbug, 156
Darwin, Charles, no place for meeting, 69
Darwin, Erasmus, 23, 153
Dazio, Signor Pietro, of Fusio, 140
Death, no man can die to himself, 277
Deceit a necessary alloy of truth, 280

Index

329

- Department, good technique resembles, 156
Desire and power, 108
Development of power to know our own likes and dislikes, 22
Devil's Bridge, 23
Diatonic scale, and song of birds in New Zealand, 232
Dirt, eating a peck of moral, 71
Disgrazia and misfortune, 58
D'Israeli, Isaac, quotations from, 67
Dissenters all narrow-minded, 153
Distribution of plants and animals often inexplicable, 135
Diversion of mental images, 54
Doera, fresco at, 145, 221
Dogs, 156, 202, 260, 313
Doing, the only mode of learning, 151
Doors, how they open in time, 151
Doubt, "There lives more doubt in honest faith," 67
Downs, the South, like Monte Generoso, 230
Draughtsman, first business of a, 148
Drawing, the old manner of teaching, 150
Dream, my, at Lago di Cadagno, 82
Drunkenness and imagination, 46
Dunque, 133
Duso, Agostino, his fresco at Sta. Maria in Calanca, 225
Earnestness, 142, 192
Eating, a mode of bigotry, 153
Echo at Graglia, 192
Edelweiss, 291
Electricity and Alpine roads, 60
Elephant brays a third, 233
"Elongated" honey, 293
Embryonic stages, the artist
- English priests and Italian, 106
— why introspective, 18
Equilibrium only attainable at the cost of progress, 195
Eritis, a panic concerning, 204
Eternal punishment, 111, 196
Eusebius, St., 178
Evolution and illusion, 43
— essence of, consists in not shocking too much, 110
Extreme, every, an absurdity, 153
Faido, 22
Faith, doubt lives in honest, 67
— more assured in the days of spiritual Saturnalia, 68
— foundations of our system based on, 107, 277
— and reason, 108
— catholic, of protoplasm, 152
— a mode of impudence, 283
Falsehood turning to truth, 71
Famine prices at Locarno, 276
Feeling, current, the safest guide, 108
Fertile, rich and poor rarely fertile *inter se*, 195
Fires, how Italians manage their, 117
Fishmonger choosing a bloater, 23
Flats and sharps, a maze of metaphysical, 23
Fleet Street, beauties of, 19
Flowers, names of, 291
Fossil-soul, 234
Foundations of action lie deeper than reason, 107
— of a durable system laid on faith, 277
Francis, St., and Insurance Co.'s plate, 191
Friction, which prevents the unduly rapid growth of inventions, 60

Inverted Index

328

Index

- Canvas of life turned upside down, 68
"Carbonate of pork," 315
Carracci, the, 147
Casina di Banda, 119
Castelletto, 265
Cavagnago, 76
Cenere, Monte, narcissuses on, 228
Ceres, 161
Cerrea, 133
Chalk, Conté, the Italian for whom this was the one thing needful, 136
Chalk eggs, 43
Chamois, foot of, 283
Change, repudiation of desire for sudden, 186
— importance of, depends on the rate of introduction, 196
— either the circumstances or the sufferer will, 196
Changes, sweeping, to be felt hereafter as vibrations, 60
Cheapissimo, 165
Cheese and the *albi*, 289
Cherries, 33, 35, 49
Chestnuts, 118
Chicory and seed onions, weary utterness in, 227
Children, subalpine, 301
— what becomes of the clever, 149
Chinese, the examination-ridden, 151
Chironico, 75
"Chow," 52
Church-going, subalpine, 303
Circulation of people like blood, 20
Ciseri, his picture at Locarno, 271
Civilisation, antiquity of Italian, 124
— stationary, of ants and bees, 195
Class distinction inevitable, 195
Classification only possible through sense of shock, 63
Clergy, our English and S. Michele

- Cocking, Wednesbury, 55, 305
Collects, unsympathetic priest bristling with, 111
Colleone, Medea, 231
Colma di San Giovanni, 163
Comba di Susa, 119
Comfort as a moral influence, 185
Comic song, the landlord's, 128
Common sense, the safest guide, 108
Consistent, who ever is? 153
Contradictory principles, there must be a harmonious fusing of, 152
Converting things by eating them, 153
Corpses, desiccated, at S. Michele, 97
Cousins, my, the lower animals, 60
Cows fighting in farmyard, 120
Cricco, 125
Cristoforo, S., church of, at Mesocco, 208
— — — — at Castello, 234
Crossing, efficacy of, 152
— unexpected results of, 55
— useless if too wide, 157
Crucifixion, fresco at Fusio, 140
Culture and priggishness, 141
— a mode of concealing weakness, 192
Current feeling, the safest guide, 108
Cutlets, burnt, and the waiter, 144
Dalpe, 38
Dante a humbug, 156
Darwin, Charles, no place for meeting, 69
Darwin, Erasmus, 23, 133
Dazio, Signor Pietro, of Fusio, 140
Death, no man can die to himself, 277
Deceit a necessary alloy of truth, 280

Index

329

- Department, good technique resembles, 156
Desire and power, 108
Development of power to know our own likes and dislikes, 22
Devil's Bridge, 23
Diatonic scale, and song of birds in New Zealand, 232
Dirt, eating a peck of moral, 71
Disgrazia and misfortune, 58
D'Israeli, Isaac, quotations from, 67
Dissenters all narrow-minded, 153
Distribution of plants and animals often inexplicable, 135
Diversion of mental images, 54
Doera, fresco at, 145, 221
Dogs, 156, 202, 260, 313
Doing, the only mode of learning, 151
Doors, how they open in time, 151
Doubt, "There lives more doubt in honest faith," 67
Downs, the South, like Monte Generoso, 230
Draughtsman, first business of a, 148
Drawing, the old manner of teaching, 150
Dream, my, at Lago di Cadagno, 82
Drunkenness and imagination, 46
Dunque, 133
Duso, Agostino, his fresco at Sta. Maria in Calanca, 225
Earnestness, 142, 192
Eating, a mode of bigotry, 153
Echo at Graglia, 192
Edelweiss, 291
Electricity and Alpine roads, 60
Elephant brays a third, 233
"Elongated" honey, 293
Embryonic stages, the artist

- English priests and Italian, 106
— why introspective, 18
Equilibrium only attainable at the cost of progress, 195
Eritis, a panic concerning, 204
Eternal punishment, 111, 196
Eusebius, St., 178
Evolution and illusion, 43
— essence of, consists in not shocking too much, 110
Extreme, every, an absurdity, 153
Faido, 22
Faith, doubt lives in honest, 67
— more assured in the days of spiritual Saturnalia, 68
— foundations of our system based on, 107, 277
— and reason, 108
— catholic, of protoplasm, 152
— a mode of impudence, 283
Falsehood turning to truth, 71
Famine prices at Locarno, 276
Feeling, current, the safest guide, 108
Fertile, rich and poor rarely fertile *inter se*, 195
Fires, how Italians manage their, 117
Fishmonger choosing a bloater, 23
Flats and sharps, a maze of metaphysical, 23
Fleet Street, beauties of, 19
Flowers, names of, 291
Fossil-soul, 234
Foundations of action lie deeper than reason, 107
— of a durable system laid on faith, 277
Francis, St., and Insurance Co.'s plate, 191
Friction, which prevents the unduly rapid growth of inventions, 60

Inverted Index

328

Index

- Canvas of life turned upside down, 68
"Carbonate of pork," 315
Carracci, the, 147
Casina di Banda, 119
Castelletto, 265
Cavagnago, 76
Cenere, Monte, narcissuses on, 228
Ceres, 161
Cerrea, 133
Chalk, Conté, the Italian for whom this was the one thing needful, 136
Chalk eggs, 43
Chamois, foot of, 283
Change, repudiation of desire for sudden, 186
— importance of, depends on the rate of introduction, 196
— either the circumstances or the sufferer will, 196
Changes, sweeping, to be felt hereafter as vibrations, 60
Cheapissimo, 165
Cheese and the *alpi*, 289
Cherries, 33, 35, 46
Chestnuts, 118
Chicory and seed onions, weary utterness in, 227
Children, subalpine, 301
— what becomes of the clever, 149
Chinese, the examination-ridden, 151
Chironico, 75
"Chow," 52
Church-going, subalpine, 303
Circulation of people like blood, 20
Ciseri, his picture at Locarno, 271
Civilisation, antiquity of Italian, 124
— stationary, of ants and bees, 195
Class distinction inevitable, 195
Classification only possible through sense of shock, 63
Clergy, our English and S. Michele
- Cocking, Wednesbury, 55, 305
Collects, unsympathetic priest bristling with, 111
Colleone, Medea, 231
Colma di San Giovanni, 163
Comba di Susa, 119
Comfort as a moral influence, 185
Comic song, the landlord's, 128
Common sense, the safest guide, 108
Consistent, who ever is? 153
Contradictory principles, there must be a harmonious fusing of, 152
Converting things by eating them, 153
Corpses, desiccated, at S. Michele, 97
Cousins, my, the lower animals, 60
Cows fighting in farmyard, 120
Cricco, 125
Cristoforo, S., church of, at Mesocco, 208
— — — — at Castello, 234
Crossing, efficacy of, 152
— unexpected results of, 55
— useless if too wide, 157
Crucifixion, fresco at Fusio, 140
Culture and priggishness, 141
— a mode of concealing weakness, 192
Current feeling, the safest guide, 108
Cutlets, burnt, and the waiter, 144
- Dalpe, 38
Dante a humbug, 156
Darwin, Charles, no place for meeting, 69
Darwin, Erasmus, 23, 133
Dazio, Signor Pietro, of Fusio, 140
Death, no man can die to himself, 277
Deceit a necessary alloy of truth, 280

Index

329

- Department, good technique resembles, 156
Desire and power, 108
Development of power to know our own likes and dislikes, 22
Devil's Bridge, 23
Diatonic scale, and song of birds in New Zealand, 232
Dirt, eating a peck of moral, 71
Disgrazia and misfortune, 58
D'Israeli, Isaac, quotations from, 67
Dissenters all narrow-minded, 153
Distribution of plants and animals often inexplicable, 135
Diversion of mental images, 54
Doera, fresco at, 145, 221
Dogs, 156, 202, 260, 313
Doing, the only mode of learning, 151
Doors, how they open in time, 151
Doubt, "There lives more doubt in honest faith," 67
Downs, the South, like Monte Generoso, 230
Draughtsman, first business of a, 148
Drawing, the old manner of teaching, 150
Dream, my, at Lago di Cadagno, 82
Drunkenness and imagination, 46
Dunque, 133
Duso, Agostino, his fresco at Sta. Maria in Calanca, 225
- Earnestness, 142, 192
Eating, a mode of bigotry, 153
Echo at Graglia, 192
Edelweiss, 291
Electricity and Alpine roads, 60
Elephant brays a third, 233
"Elongated" honey, 293
Embryonic stages, the artist
- English priests and Italian, 106
— why introspective, 18
Equilibrium only attainable at the cost of progress, 195
Eritis, a panic concerning, 204
Eternal punishment, 111, 196
Eusebius, St., 178
Evolution and illusion, 43
— essence of, consists in not shocking too much, 110
Extreme, every, an absurdity, 153
- Faido, 22
Faith, doubt lives in honest, 67
— more assured in the days of spiritual Saturnalia, 68
— foundations of our system based on, 107, 277
— and reason, 108
— catholic, of protoplasm, 152
— a mode of impudence, 283
Falsehood turning to truth, 71
Famine prices at Locarno, 276
Feeling, current, the safest guide, 108
Fertile, rich and poor rarely fertile *inter se*, 195
Fires, how Italians manage their, 117
Fishmonger choosing a bloater, 23
Flats and sharps, a maze of metaphysical, 23
Fleet Street, beauties of, 19
Flowers, names of, 291
Fossil-soul, 234
Foundations of action lie deeper than reason, 107
— of a durable system laid on faith, 277
Francis, St., and Insurance Co.'s plate, 191
Friction, which prevents the unduly rapid growth of inventions, 60

About Apache Lucene

Library behind search servers

Elasticsearch + **Apache Solr**

Users?



WIKIPEDIA
The Free Encyclopedia

Users?



WIKIPEDIA
The Free Encyclopedia



GitHub

Users?



WIKIPEDIA
The Free Encyclopedia



GitHub

Users?





WIKIPEDIA
The Free Encyclopedia



GitHub



users?

Apple iTunes

THE APACHE™ SOFTWARE FOUNDATION
<http://www.apache.org/>

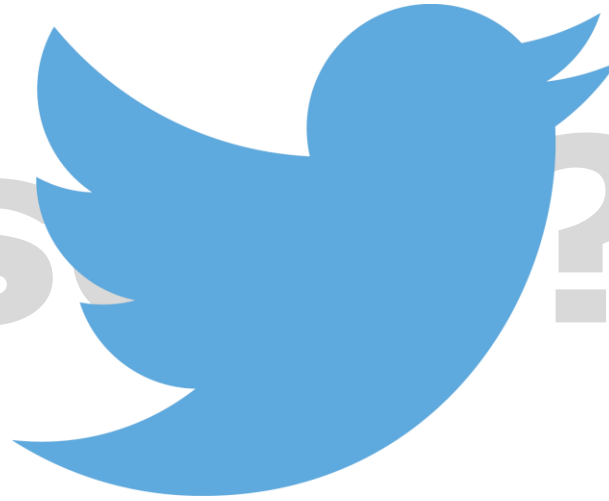
Lucene



WIKIPEDIA
The Free Encyclopedia



GitHub



Apple iTunes



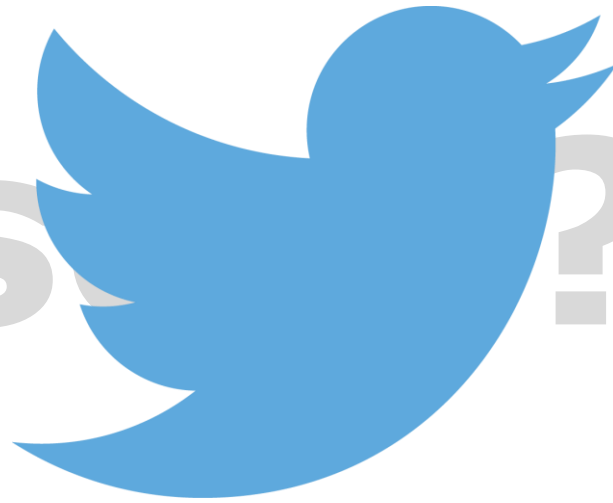
WIKIPEDIA
The Free Encyclopedia



GitHub



Apple iTunes



JIRA

Apache Lucene

ALGORITHMS ???

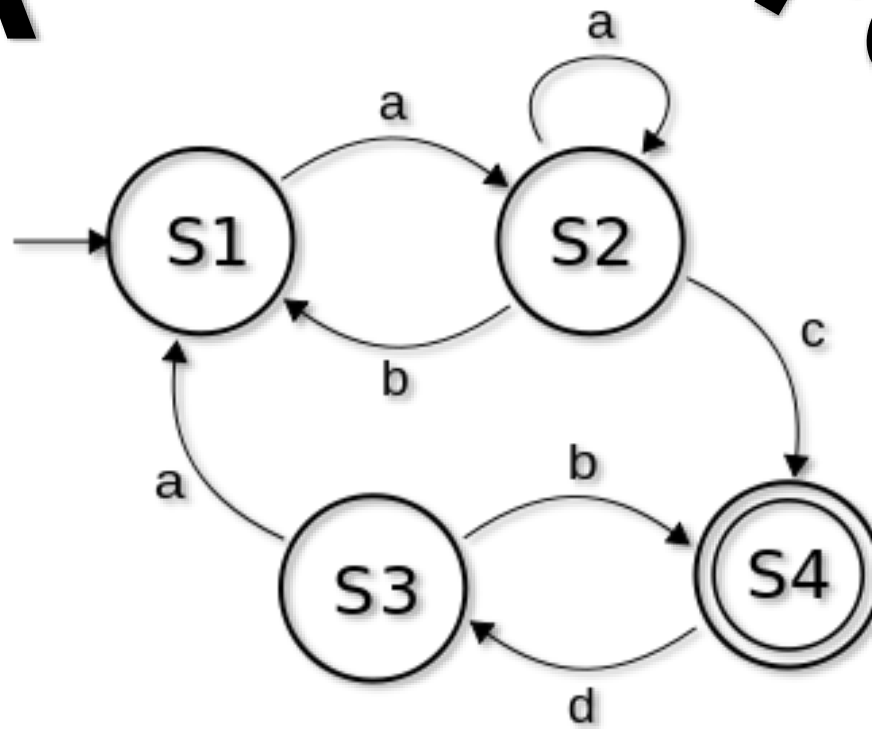

```
learnCollector, singleClauseCollector, bits acceptDocs, int
final int windowBase = top.next & ~MASK; // find the window
final int windowMin = Math.max(min, windowBase);
final int windowMax = Math.min(max, windowBase + SIZE);

// Fill 'leads' with all scorers from 'head' that are in the
leads[0] = head.pop();
int maxFreq = 1;
while (head.size() > 0 && head.pop().get() < windowMax) {
    leads[maxFreq++] = head.pop();
}

if (minThroughMax == 1 && maxFreq == 1) {
    // special case: only a single scorer can match in the current
    // we can collect directly
    final BulkScorerAndDocBulkScorer = leads[0];
    scorer = new SingleScorer(bulkScorer, singleClauseCollector,
```

**Performance critical,
tight loops**

FSA



FST



Everywhere!

```
@Override
public int nextDoc() throws IOException {
    if (docUpto == docFreq) {
        return doc = NO_MORE_DOCS;
    }
    if (docBufferUpto == BLOCK_SIZE) {
        refillDocs();
    }

    accum += docDeltaBuffer[docBufferUpto];
    freq = freqBuffer[docBufferUpto];
    posPendingCount += freq;
    docBufferUpto++;
    docUpto++;

    doc = accum;
    position = 0;
    lastStartOffset = 0;
    return doc;
}

@Override
public int advance(int target) throws IOException {
    // TODO: make frq block load lazy/skipable

    if (target > nextSkipDoc) {
        if (skipper == null) {
            // Lazy init: first time this enum has ever been used for skipping
            skipper = new Lucene50SkipReader(docIn.clone(),
                MAX_SKIP_LEVELS,
                true,
                indexHasOffsets,
                indexHasPayloads);
        }

        if (!skipped) {
            assert skipOffset != -1;
            // This is the first time this enum has skipped
        }
    }
}
```

Iterators!
Iterators!
Iterators!



```
@Override
public int nextDoc() throws IOException {
    if (docUpto == docFreq) {
        return doc = NO_MORE_DOCS;
    }
    if (docBufferUpto == BLOCK_SIZE) {
        refillDocs();
    }

    accum += docDeltaBuffer[docBufferUpto];
    freq = freqBuffer[docBufferUpto];
    posPendingCount += freq;
    docBufferUpto++;
    docUpto++;

    doc = accum;
    position = 0;
    lastStartOffset = 0;
    return doc;
}

@Override
public int advance(int target) throws IOException {
    // TODO: make frq block load lazy/skipable

    if (target > nextSkipDoc) {
        if (skipper == null) {
            // Lazy init: first time this enum has ever been used for skipping
            skipper = new Lucene50SkipReader(docIn.clone(),
                MAX_SKIP_LEVELS,
                true,
                indexHasOffsets,
                indexHasPayloads);
        }

        if (!skipped) {
            assert skipOffset != -1;
            // This is the first time this enum has skipped
        }
    }
}
```

Iterators!
Iterators!
Iterators!

Intersection
while iterating!



Bitset operations (AND, OR, nextSetBit)

Rare use of String Slices on byte[]

The Generics Policeman Blog

[My Homepage / Legal Notice](#)

[Apache Lucene](#)

[SD DataSolutions GmbH](#)

[Schindlers Software](#)

[PANGAEA](#)



2012-07-18

Use Lucene's MMapDirectory on 64bit systems, please!

Don't be afraid – Some warfar can't come in misunderstandings

Since version 3.1, Apache Lucene and Solr use MMapDirectory by default on 64bit Windows and Solaris systems; since version 3.3 also for 64bit Linux systems. This change lead to some confusion among Lucene and Solr users, because suddenly their systems started to behave differently than in previous versions. On the Lucene and Solr mailing lists a lot of posts arrived from users asking why their Java installation is suddenly consuming three times their physical memory or system administrators complaining about heavy resource usage. Also consultants were starting to tell people that they should **not** use MMapDirectory and change their solrconfig.xml to work instead with slow SimpleFSDirectory or NIOFSDirectory (which is much slower on Windows, caused by a JVM bug #6265734). From the point of view of the Lucene committers, who carefully decided that using MMapDirectory is the best for those platforms, this is rather annoying, because they know, that Lucene/Solr can work with much better performance than before. Common misinformation about the background of this change causes suboptimal installations of this great search engine everywhere.

In this blog post, I will try to explain the basic operating system facts regarding virtual memory handling in the kernel and how this can be used to largely improve performance of Lucene

About me



Uwe Schindler

Follow

243

I work as committer and PMC member for Apache Lucene and PANGAEA. I invented the numeric range functionality for fast range queries (NumericRangeQuery). Currently I also have the position of the project chair. I am also working in the PHP development crew, maintaining the web server plug-in for Sun Java System Web Servers (my favourite web server). My (IT) interests are Lucene Java, XML techniques, Data Warehousing, Sensor Networks, metadata dissemination using global standards, global unique identifiers like DOIs,... The recently founded "SD DataSolutions GmbH" offers support for Lucene-related search solutions, especially numerical searches and other advanced indexing techniques.

[View my complete profile](#)





Dawid Weiss: More Challenges for JVM!

**RANDOMIZE YOUR TESTS
AND IT WILL BLOW YOUR
SOCKS OFF! *)**

Randomization everywhere

- Input data, iteration counts, arguments.
 - Random, constraint-bound, shuffled
- Software components.
 - If multiple implementations exist: Field, Directory abstraction, IndexSearcher...
- Environment.
 - Locale, Timezone,...
 - *JVM (!), operating system*
- Exceptional triggers.
 - I/O problems, network problems (using mocks or runtime engineering)

RandomizedRunner's goals

Compatibility

with JUnit (and tools). At 99%, relax contracts when useful.

Built-in randomization

including reporting/ stack augmentations.

Test isolation

by tracking spawned threads. Timeouts. Terminations.

Utilities

@Repeat, @Seed, @Nightly, @TestGroup, @TestFactories...

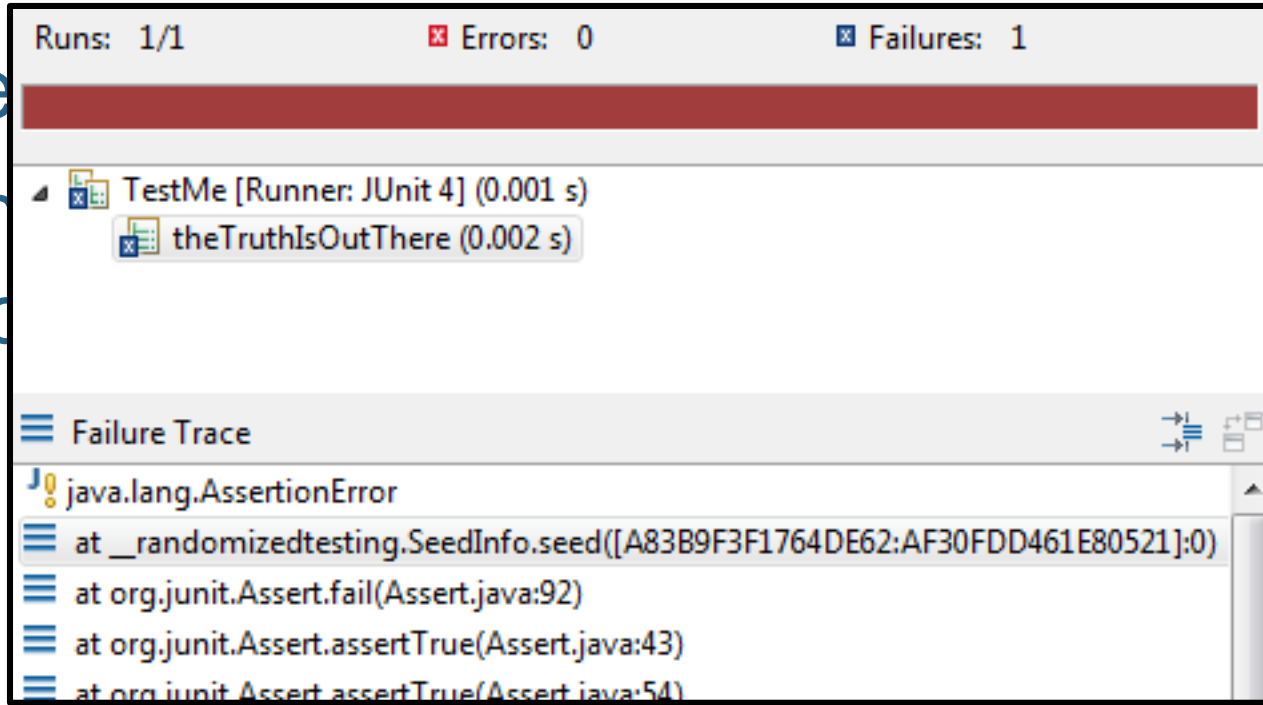
<https://github.com/randomizedtesting/randomizedtesting>

Reproducibility?

- Every test gets an initial random seed
- Printed on test execution & included in stack traces

Reproducibility?

- Even
- Print
- Stack



The screenshot shows a test runner window with the following details:

- Runs: 1/1
- Errors: 0
- Failures: 1

The test results section shows:

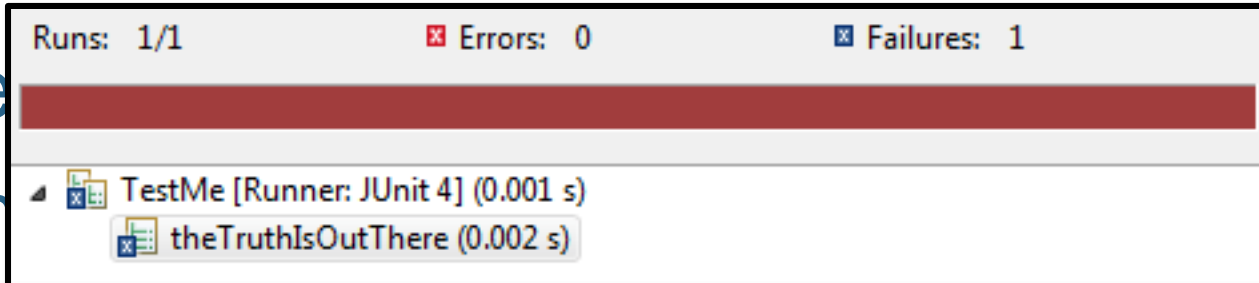
- TestMe [Runner: JUnit 4] (0.001 s)
 - theTruthIsOutThere (0.002 s)

The Failure Trace section shows the following stack trace:

```
java.lang.AssertionError
at __randomizedtesting.SeedInfo.seed([A83B9F3F1764DE62:AF30FDD461E80521]:0)
at org.junit.Assert.fail(Assert.java:92)
at org.junit.Assert.assertTrue(Assert.java:43)
at org.junit.Assert.assertTrue(Assert.java:54)
```

Reproducibility?

- Even
- Print



In case the JVM freezes, use a signal or jstack to get the master seed.

```
> jstack 6502
```

```
...  
"TEST-TestScope-TestHanging.myTest-seed#[96F549A92265CBF8]"  
waiting on condition [1169ef000]  
  java.lang.Thread.State: TIMED_WAITING (sleeping)  
  at java.lang.Thread.sleep(Native Method)  
  at TestHanging.myTest(TestHanging.java:26)
```

Assertions in randomized test code?

- Compare against reference.
 - Naïve, previous or alternative implementations.
- Sanity checks.
 - Crude output checks (boundary conditions).
 - Sanity assertions inside code.
- Nothing!
 - Unchecked exceptions. **Or a JVM core dump.**
Surprisingly effective :)



24/7 randomized testing using many JVMs (-settings)

POLICEMAN JENKINS

Missing parts

- JVM randomization
 - Oracle JDK 7, Oracle JDK 8
 - IBM J9
 - Preview releases: JDK 9 EA

Missing parts

- JVM randomization
 - Oracle JDK 7, Oracle JDK 8
 - IBM J9
 - Preview releases: JDK 9 EA
- JVM settings randomization
 - Garbage collector
 - Bitness: 32 / 64 bits
 - Server / Client VM
 - Compressed OOPs (*ordinary object pointer*)

Missing parts

- JVM randomization
 - Oracle JDK 7, Oracle JDK 8
 - IBM J9
 - Preview releases: JDK 9 EA
- JVM settings randomization
 - Garbage collector
 - Bitness: 32 / 64 bits
 - Server / Client VM
 - Compressed OOPs (*ordinary object pointer*)
- Platform
 - Linux, Windows, MacOS X, Solaris

Dashboard [Jenkins] x jenkins.thetaphi.de

Uwe Schindler | log out

Policeman Jenkins search

Jenkins ENABLE AUTO REFRESH

- New Item
- People
- Build History
- Manage Jenkins
- Credentials
- My Views
- Job Config History
- We Need Beer

Build Queue

No builds in the queue.

Build Executor Status

- master
 - 1 [Lucene-Solr-trunk-Linux](#) #15675
- MacOSX VBOX
 - 1 Idle
- Solaris VBOX
 - 1 Idle
- Windows VBOX
 - 1 [Lucene-Solr-trunk-Windows](#) #5575

Policeman Jenkins Server – monitoring [Apache Lucene](#) and [Apache Solr](#)

This server is provided by [SD DataSolutions GmbH](#). It supports testing of the code base with various JDK versions, provided by Oracle. It checks the whole range of versions from Oracle JDK 7 to JDK 8 and preview builds of Java 9. Each run uses different settings, like bitness, garbage collector, or reference size. Support for Windows, Mac OS X, and Solaris is provided by [Oracle VirtualBox \(admin\)](#).

[edit description](#)

All +

S	W	Name ↓	Last Success	Last Failure	Last Duration
		Forbidden-APIs	4 days 20 hr - #312	N/A	57 sec
		Forbidden-APIs-testonly-MacOSX	4 days 19 hr - #6	1 mo 0 days - #1	1 min 47 sec
		Lucene-Solr-5.x-Linux	1 hr 26 min - #15372	3 hr 33 min - #15371	1 hr 3 min
		Lucene-Solr-5.x-MacOSX	13 hr - #2986	5 hr 49 min - #2987	1 hr 49 min
		Lucene-Solr-5.x-Solaris	2 hr 59 min - #354	19 hr - #353	1 hr 33 min
		Lucene-Solr-5.x-Windows	9 hr 8 min - #5442	1 day 11 hr - #5439	1 hr 40 min
		Lucene-Solr-trunk-Linux	2 hr 33 min - #15674	6 hr 32 min - #15672	1 hr 7 min
		Lucene-Solr-trunk-MacOSX	7 hr 28 min - #3042	1 day 9 hr - #3039	1 hr 38 min
		Lucene-Solr-trunk-Solaris	4 hr 34 min - #360	20 hr - #358	1 hr 34 min
		Lucene-Solr-trunk-Windows	10 hr - #5574	17 hr - #5573	1 hr 39 min

Icon: [S](#) [M](#) [L](#)

Legend [RSS for all](#) [RSS for failures](#) [RSS for just latest builds](#)

- Back to Dashboard
- Status
- Changes
- Workspace
- Build Now
- Delete Project
- Configure
- Email Template Testing
- Java Warnings
- GroovyScriptTrigger Log
- Job Config History

Project Lucene-Solr-trunk-Linux

[add description](#)
 Disable Project

"All my life, I always wanted to be somebody. Now I see that I should have been more specific." Jane Wagner

- Workspace
- Recent Changes
- Latest Test Result (no failures)

Upstream Projects

[Lucene-Solr-5.x-Linux](#)

Downstream Projects

[Lucene-Solr-5.x-Linux](#)

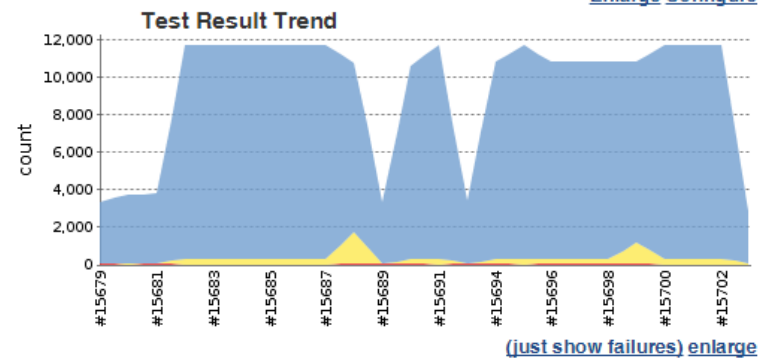
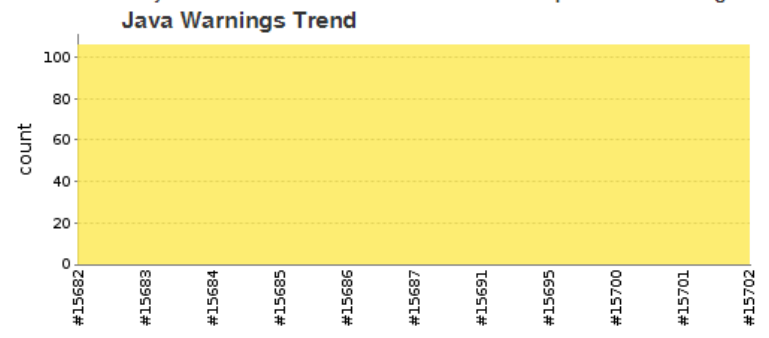
Permalinks

- [Last build \(#15704\), 1 min 53 sec ago](#)
- [Last stable build \(#15702\), 3 hr 4 min ago](#)
- [Last successful build \(#15702\), 3 hr 4 min ago](#)
- [Last failed build \(#15703\), 1 hr 3 min ago](#)
- [Last unsuccessful build \(#15703\), 1 hr 3 min ago](#)
- [Last completed build \(#15703\), 1 hr 3 min ago](#)

Build History [trend](#)

find

- #15704
Jan 30, 2016 10:34 AM
Java: 32bit/jdk-1.8.0_72 -client -XX:+UseParallelGC
- #15703
Jan 30, 2016 9:32 AM
Java: 32bit/jdk-9-ea+102 -server -XX:+UseG1GC -XX:-CompactStrings
- #15702
Jan 30, 2016 7:31 AM
Java: 64bit/jdk-9-ea+102 -XX:+UseCompressedOops -XX:+UseG1GC -XX:-CompactStrings
- #15701
Jan 30, 2016 5:44 AM
Java: 64bit/jdk-9-ea+102 -XX:-UseCompressedOops -XX:+UseConcMarkSweepGC -XX:-CompactStrings



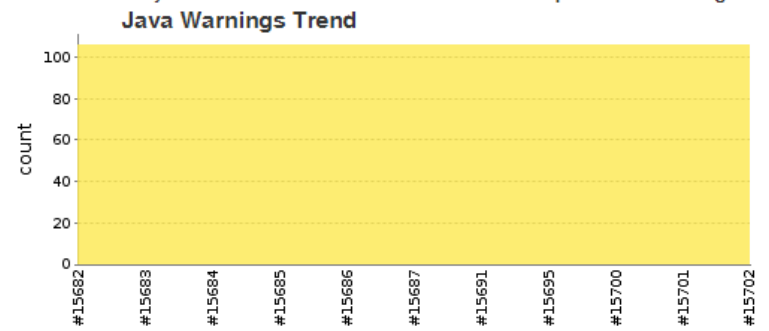
- Back to Dashboard
- Status
- Changes
- Workspace
- Build Now
- Delete Project
- Configure
- Email Template Testing
- Java Warnings
- GroovyScriptTrigger Log
- Job Config History

Project Lucene-Solr-trunk-Linux

[add description](#)
[Disable Project](#)

"All my life, I always wanted to be somebody. Now I see that I should have been more specific." Jane Wagner

- Workspace
- Recent Changes
- Latest Test Result (no failures)



Upstream Projects

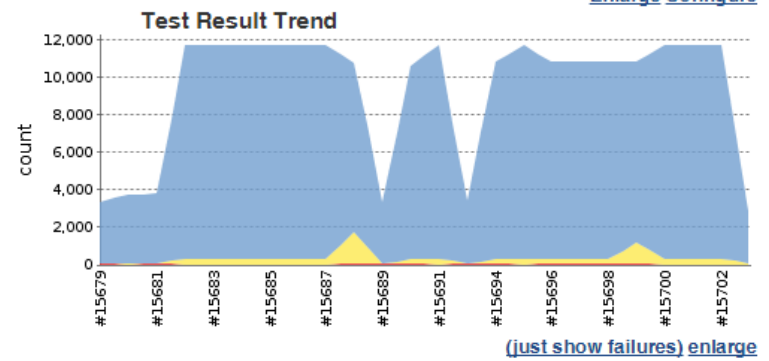
[Lucene-Solr-5.x-Linux](#)

Downstream Projects

[Lucene-Solr-5.x-Linux](#)

Permalinks

- [Last build \(#15704\), 1 min 53 sec ago](#)
- [Last stable build \(#15702\), 3 hr 4 min ago](#)
- [Last successful build \(#15702\), 3 hr 4 min ago](#)
- [Last failed build \(#15703\), 1 hr 3 min ago](#)
- [Last unsuccessful build \(#15703\), 1 hr 3 min ago](#)
- [Last completed build \(#15703\), 1 hr 3 min ago](#)



Build History [trend](#)

find

- #15704**
Jan 30, 2016 10:34 AM
Java: 32bit/jdk1.8.0_72 -client -XX:+UseParallelGC
- #15703**
Jan 30, 2016 9:32 AM
Java: 32bit/jdk-9-ea+102 -server -XX:+UseG1GC -XX:-CompactStrings
- #15702**
Jan 30, 2016 7:31 AM
Java: 64bit/jdk-9-ea+102 -XX:+UseCompressedOops -XX:+UseG1GC -XX:-CompactStrings
- #15701**
Jan 30, 2016 5:44 AM
Java: 64bit/jdk-9-ea+102 -XX:-UseCompressedOops -XX:+UseConcMarkSweepGC -XX:-CompactStrings

Uwe

Lucene-Solr-trunk-Linux [X]

jenkins.thetaphi.de/job/Lucene-Solr-trunk-Linux/

Jenkins > Lucene-Solr-trunk-Linux > ENABLE AUTO REFRESH

[Back to Dashboard](#)

Project Lucene-Solr-trunk-Linux

[Status](#)

```
[script-trigger] Groovy expression evaluation to true. (log)
[EnvInject] - Loading node environment variables.
[EnvInject] - Preparing an environment for the build.
[EnvInject] - Keeping Jenkins system variables.
[EnvInject] - Keeping Jenkins build variables.
[EnvInject] - Evaluation the following Groovy script content:
return evaluate(new java.io.File(JENKINS_HOME, "scripts/linux-random-java8.groovy"))

Using Java: 32bit/jdk-9-ea+102 -server -XX:+UseG1GC -XX:-CompactStrings
[EnvInject] - Injecting contributions.
Building on master in workspace /var/lib/jenkins/workspace/Lucene-Solr-trunk-Linux
Fetching changes from the remote Git repository
Cleaning workspace
Checking out Revision 9332b1602cc0f7312fc22a3d088c549299015691 (refs/remotes/origin/master)
No emails were triggered.
[description-setter] Description set: Java: 32bit/jdk-9-ea+102 -server -XX:+UseG1GC -XX:-CompactStrings
[Lucene-Solr-trunk-Linux] $ /var/lib/jenkins/tools/hudson.tasks.Ant_AntInstallation/ANT_1.8.2/bin/ant "-Dargs=-server -XX:+UseG1GC -XX:-CompactStrings" jenkins-hourly
Buildfile: /home/jenkins/workspace/Lucene-Solr-trunk-Linux/build.xml

jenkins-hourly:

-print-java-info:
[java-info] java version "9-ea"
[java-info] Java(TM) SE Runtime Environment (9-ea+102-2016-01-21-001243.javare.4316.nc, Oracle Corporation)
[java-info] Java HotSpot(TM) Server VM (9-ea+102-2016-01-21-001243.javare.4316.nc, Oracle Corporation)
[java-info] Test args: [-server -XX:+UseG1GC -XX:-CompactStrings]

clean:
-XX:+UseG1GC -XX:-CompactStrings
```

- [Last unsuccessful build \(#15703\), 1 hr 3 min ago](#)
- [Last completed build \(#15703\), 1 hr 3 min ago](#)

#15701

Jan 30, 2016 5:44 AM

Java: 64bit/jdk-9-ea+102 -XX:-UseCompressedOops
-XX:+UseConcMarkSweepGC -XX:-CompactStrings

Lucene-S

Jenkins

Back

Status

find

#15701

Jan 30,

Java: 64

-XX:+Us

```
[junit4]
[junit4] Suite: org.apache.lucene.search.TestSimilarityProvider
[junit4] Completed [25/410] on J1 in 0.05s, 1 test
[junit4]
[junit4] JVM J0: stdout was not empty, see: /home/jenkins/workspace/Lucene-Solr-trunk-
Linux/lucene/build/core/test/temp/junit4-J0-20160130_093254_626.sysout
[junit4] >>> JVM J0 emitted unexpected output (verbatim) ----
[junit4] #
[junit4] # A fatal error has been detected by the Java Runtime Environment:
[junit4] #
[junit4] # SIGSEGV (0xb) at pc=0xef297ffb, pid=18795, tid=19014
[junit4] #
[junit4] # JRE version: Java(TM) SE Runtime Environment (9.0+102) (build 9-ea+102-2016-01-21-
001243.javare.4316.nc)
[junit4] # Java VM: Java HotSpot(TM) Server VM (9-ea+102-2016-01-21-001243.javare.4316.nc,
mixed mode, tiered, g1 gc, linux-x86)
[junit4] # Problematic frame:
[junit4] # J 5236 C2
org.apache.lucene.codecs.compressing.CompressingTermVectorsWriter.flushOffsets([I)V (820 bytes)
@ 0xef297ffb [0xef297d20+0x000002db]
[junit4] #
[junit4] # No core dump will be written. Core dumps have been disabled. To enable core
dumping, try "ulimit -c unlimited" before starting Java again
[junit4] #
[junit4] # An error report file with more information is saved as:
[junit4] # /home/jenkins/workspace/Lucene-Solr-trunk-
Linux/lucene/build/core/test/J0/hs_err_pid18795.log
[junit4] Could not load hsdisk-i386.so; library not loadable; PrintAssembly is disabled
[junit4] [thread 19028 also had an error]
[junit4] [thread 19016 also had an error]
[junit4] #
[junit4] # If you would like to submit a bug report, please visit:
[junit4] # http://bugreport.java.com/bugreport/crash.jsp
[junit4] #
[junit4] <<< JVM J0: EOF ----
[junit4] Suite: org.apache.lucene.index.TestIndexWriterMerging
[junit4] Completed [26/410] on J2 in 3.89s, 6 tests
[junit4]
```

ESH

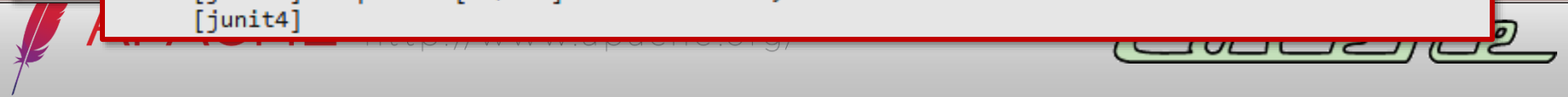
tion

er

#15702

re

ge



Testing JDK

BUGS FOUND

- **Java 7 GA**

- *let's don't talk about it!*





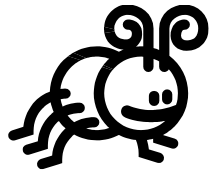
- **Java 7 GA**
 - *let's don't talk about it!*
- **Java 7u40: AVX optimizations broken (JDK-8024830)**
 - Haswell or later CPU
 - Fixed in 7u55



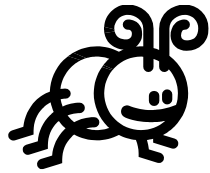
- **Java 7 GA**
 - *let's don't talk about it!*
- **Java 7u40: AVX optimizations broken (JDK-8024830)**
 - Haswell or later CPU
 - Fixed in 7u55
- **Java 7 / 8: Runtime.exec() fails in Turkish locale (JDK-8047340)**
 - Fixed in 8u40



- **Java 7 GA**
 - *let's don't talk about it!*
- **Java 7u40: AVX optimizations broken (JDK-8024830)**
 - Haswell or later CPU
 - Fixed in 7u55
- **Java 7 / 8: Runtime.exec() fails in Turkish locale (JDK-8047340)**
 - Fixed in 8u40
- **Java 7u25: ByteSliceReader (Lucene class) assert trips with 32-bit 7u25 + G1GC (JDK-8038348)**
 - Hard to reproduce, cause still unknown!

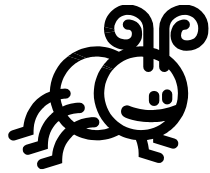


Java 9 Bug Parade



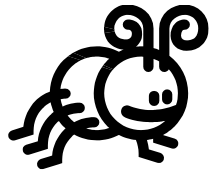
Java 9 Bug Parade

- Array-Copy bugs (*JDK-8134468*, *JDK-8080976*,...)
 - easy to reproduce



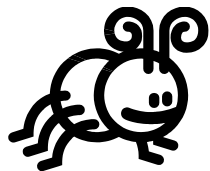
Java 9 Bug Parade

- Array-Copy bugs (*JDK-8134468, JDK-8080976, ...*)
 - easy to reproduce
- String.toLowerCase do not work for some concatenated strings (*JDK-8042589*)
 - another Hotspot issue



Java 9 Bug Parade

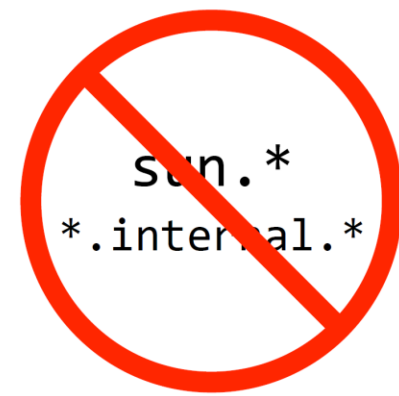
- Array-Copy bugs (*JDK-8134468, JDK-8080976, ...*)
 - easy to reproduce
- `String.toLowerCase` do not work for some concatenated strings (*JDK-8042589*)
 - another Hotspot issue
- **JDK 9 b93** breaks Apache Lucene due to compact strings (*JDK-8144212*)
 - easy to reproduce
 - fixed recently (`String#getChars()` optimization)



Java 9 Bug Parade

- Array-Copy bugs (*JDK-8134468, JDK-8080976, ...*)
 - easy to reproduce
- `String.toLowerCase` do not work for some concatenated strings (*JDK-8042589*)
 - another Hotspot issue
- **JDK 9 b93** breaks Apache Lucene due to compact strings (*JDK-8144212*)
 - easy to reproduce
 - fixed recently (`String#getChars()` optimization)
- **JDK 9 b54** breaks compiling code with source/target 1.7 and diamond operator (*JDK-8075793*)
 - bug in type system

Java 9 Jigsaw



- **Lucene fixes:**

- Removal of `AccessibleObject#setAccessible` (where possible)

- **Recent discussions:**

- `sun.misc.Cleaner` removal**

- Would be disaster for Lucene without more fixes around `MappedByteBuffer` unmapping!!!
- “workaround” available...

Thank You!

especially:

Vladimir Kozlov, Roland Westrelin,
Tobias Hartmann, Alan Bateman,
Andrew Haley, Chris Hegarty,
Rory O'Donnell

and

Mark Reinhold